CODEBOOK

INTRODUCTION:
This codebook is split into two sections. The first section, Codebook Task 1, pertains to the Tidy1.txt dataset located in the project 1 repository. The second section, Codebook Task 2, pertains the datasets Tidy2.txt, Tidy2_a & Tidy2_b, located in the project 1 repository.

Code Book Task 1.

Variables:

<u>**Variable**</u>	**Class**	**Description**
Subject	**Numeric**	**Explains which subject was giving the resulting output**
Activity	**Character**	**The output variable that the different x variables were giving that correlated to a certain movement type 1-6. We transformed these numbers into actual names for the activities (walking, walking upstairs, walking downstairs, sitting, standing, lying)**
tBodyAcc-mean()-X tBodyAcc-mean()-Y tBodyAcc-mean()-Z tBodyAcc-std()-X tBodyAcc-std()-Y tBodyAcc-std()-Z tGravityAcc-mean()-X tGravityAcc-mean()-Y tGravityAcc-mean()-Z tGravityAcc-std()-X tGravityAcc-std()-Y tGravityAcc-std()-Z tBodyAccJerk-mean()-X tBodyAccJerk-mean()-Y tBodyAccJerk-mean()-Z	**Numeric**	**These variables were given to us in the project but were not transformed in anyway. They are the means of different movement parameters in 3 dimensions (x,y,z) as well as the standard deviations of different movement parameters in 3 dimension (x,y,z).**

tBodyAccJerk-std()-X **tBodyAccJerk-std()-Y** **tBodyAccJerk-std()-Z** **tBodyGyro-mean()-X** **tBodyGyro-mean()-Y** **tBodyGyro-mean()-Z** **tBodyGyro-std()-X** **tBodyGyro-std()-Y** **tBodyGyro-std()-Z** **tBodyGyroJerk-mean()-X** **tBodyGyroJerk-mean()-Y** **tBodyGyroJerk-mean()-Z** **tBodyGyroJerk-std()-X** **tBodyGyroJerk-std()-Y** **tBodyGyroJerk-std()-Z** **tBodyAccMag-mean()** **tBodyAccMag-std()** **tGravityAccMag-mean()** **tGravityAccMag-std()** **tBodyAccJerkMag-mean()** **tBodyAccJerkMag-std()** **tBodyGyroMag-mean()** **tBodyGyroMag-std()** **tBodyGyroJerkMag-mean()** **tBodyGyroJerkMag-std()** **fBodyAcc-mean()-X** **fBodyAcc-mean()-Y** **fBodyAcc-mean()-Z** **fBodyAcc-std()-X** **fBodyAcc-std()-Y** **fBodyAcc-std()-Z** **fBodyAcc-meanFreq()-X** **fBodyAcc-meanFreq()-Y** **fBodyAcc-meanFreq()-Z** **fBodyAccJerk-mean()-X** **fBodyAccJerk-mean()-Y** **fBodyAccJerk-mean()-Z** **fBodyAccJerk-std()-X** **fBodyAccJerk-std()-Y** **fBodyAccJerk-std()-Z** **fBodyAccJerk-meanFreq()-X**		

**fBodyAccJerk-meanFreq()-Y
fBodyAccJerk-meanFreq()-Z
fBodyGyro-mean()-X
fBodyGyro-mean()-Y
fBodyGyro-mean()-Z
fBodyGyro-std()-X
fBodyGyro-std()-Y
fBodyGyro-std()-Z
fBodyGyro-meanFreq()-X
fBodyGyro-meanFreq()-Y
fBodyGyro-meanFreq()-Z
fBodyAccMag-mean()
fBodyAccMag-std()
fBodyAccMag-meanFreq()
fBodyBodyAccJerkMag-mean()
fBodyBodyAccJerkMag-std()
fBodyBodyAccJerkMag-meanFreq()
fBodyBodyGyroMag-mean()
fBodyBodyGyroMag-std()
fBodyBodyGyroMag-meanFreq()
fBodyBodyGyroJerkMag-mean()
fBodyBodyGyroJerkMag-std()
fBodyBodyGyroJerkMag-meanFreq()**		

Task One Process

1) **Merge the following txt files in Excel: “X_Test.txt”, “X_train.txt”, “y_test.txt”, “y_train.txt”, “subject_test.txt”, “subject_train.txt”, and “features.txt”. Save this file as a csv and import into R Studio.**

2) Use the grep function to filter/select all of the variable names in the data headers that related to mean and standard deviation. Individually isolate “Subject”, “Activity”, “mean”, “std”, and then combine them with the ‘select’ function to create the data table “Raw_Selected”.
3) Use the ‘arrange’ function to order the subjects in ascending order.
4) Use group_by function to select “Subject” and “Activity” variables to summarize the average results of the variables per the subject’s activity type.
5) Rename Activity Variables (1-6) to corresponding activity name (“WALKING”, WALKING_UPSTAIRS”, etc). Select the Activity row from the data set using the ‘$’ notation, and then set the numeric value equal to the corresponding activity names.
6) Write final table with the ‘write.table’ function to save as txt file “tidy1.txt”.

Code Book Task 2.

Variable	Class	Description
Plant ID	Character	Each Plant ID is matched with a specific plant for each given year.
Year	Character	The year variable tells you the time period in which the data corresponds to. (Year is only the last two digits, ie. 1989 = 89).
Electricity	Character	The electricity variable describes the amount of electricity that was generated by a specific plant at a given year.
Short Tons S02	Character	The short tons SO2 variable describes the amount of emissions, specifically SO2 emissions, were produced by a specific plant at each given year.

Short Tons N0x	Character	The short tons NOx variable describes the amount of emissions, specifically NOx emissions, were produced by a specific plant at each given year.
Capital Stock	Character	The capital stock variable is described by using the net investment of each individual plant and dividing it by the Handy-Whitman Index, which is the conversion of historical cost of plant values to constant dollar values.
Employees	Character	The employees variable shows the amount of employees at each given plant at a given year.
Heat Content of Coal	Character	The heat content of coal variable describes the heat content of coal used at each given plant.
Heat Content of Oil	Character	The heat content of oil variable describes the heat content of coal used at each given plant.
Heat Content of Gas	Numeric	The heat content of gas variable describes the heat content of coal used at each given plant.
Short Tons SO2 (MwH)	Numeric	The short tons SO2 (MwH) variable describes the amount of emissions, specifically SO2 emissions, were produced by a specific plant at each given year, however this one differs from the original short tons SO2 variable

		because it is no longer in KwH units, rather it is converted into MwH.
Short Tons NOx (MwH)	Numeric	The short tons NOx (MwH) variable describes the amount of emissions, specifically NOx emissions, were produced by a specific plant at each given year, however this one differs from the original short tons NOx variable because it is no longer in KwH units, rather it is converted into MwH.
Capital Stock in 2017 Dollars	Numeric	This variable describes the capital stock for each individual plant at a given year, but it converts the original values from the capital stock variable into 2017 dollars.
Electricity - Daily Average	Numeric	This divides the electricity values by 365 in order to reach an estimated daily average of electricity generated by each plant for a given year.
Heat Content of Coal - Daily Average	Numeric	This divides the heat content of coal (in MwH) by 365 to give you an estimated average of the amount of heat content of coal was used at each plant for a given year.
Heat Content of Oil - Daily Average	Numeric	This divides the heat content of oil (in MwH) by 365 to give you an estimated average of the amount of heat content of oil was used at each plant for a given year.

Heat Content of Gas - Daily Average	**Numeric**	**This divides the heat content of gas (in MwH) by 365 to give you an estimated average of the amount of heat content of gas was used at each plant for a given year.**
Dummy	**Numeric**	**The dummy variable is introduced to split the data at the year 1990 (1990 being the cutoff), where the regulation began. From 1990 and earlier the dummy variable is designated a value of 1. From 1991 and on the dummy variable is designated a value of 0.**

Task Two Process

1) **Use read.table function to open “Panel_8595.Txt” in R.**
2) **Nullify or remove column V2, which only contained periods, and rows 1 & 2 which were empty or contained irrelevant values to the data.**
3) **Change the “V” headers set in R to their designated headers in order to accurately represent the values in each column.**
4) **Save the changes under a new panel: Pan_a**
5) **Clean the data within each column so that it can be read correctly, specifically removing spaces between numbers and at times the numbers after the spaces.**
6) **Convert values under relevant columns (ie. Short Tons SO2, Short Tons NOx, etc.) into different units (KwH to MwH for example) changing the class of the values to numeric.**
7) **Gathered daily averages of the values that were relevant to daily use (Electricity, Short Tons SO2, Short Tons NOx, etc.) by dividing the values by the number of days in the year.**
8) **Save the changes under a new panel: Pan_b**
9) **Create a dummy variable (0,1) to represent a before and after breakdown of when a policy was implemented. Dummy variable separated years up to 1990 at a value of 1, and 1991 on at a value of 0.**
10) **Save the changes under a new panel: Pan_c**
11) **Remove superfluous variables (F1 and F2) by nullifying the columns.**
12) **Convert capital stock values into 2017 dollars using BLS conversion rate of Jan’18 and Jan’73.**
13) **Remove the first two rows (rows 1 and 2) to rid the data of empty cells and NA cells.**

14) **Save the changes under a new panel: Pan_d**
15) **Save Pan_d as tidy2.txt**
16) **Open “Tidy2.Txt”. Use aggregate function to calculate the mean of all variables across all 92 power plants. Save as a new dataset “Tidy2_b.Txt”.**
17) **Open “Tidy2.Txt”. Create 16 unique subsets of the Tidy2 dataset, using the aggregate function, containing the sum of each variable by year.**
18) **Merge the unique subsets of the data into one dataset using the merge function. Save as “Tidy2_b.Txt”**

Code Tas